

Linear regression for numeric symbolic variables: an ordinary least squares approach based on Wasserstein Distance

Antonio Irpino, Rosanna Verde

Dipartimento di Studi Europei e Mediterranei, Seconda Universit degli Studi di Napoli, Via del Setificio, 81100, Caserta, Italy

Abstract

In this paper we present a linear regression model for modal symbolic data. The observed variables are histogram variables according to the definition given in Bock and Diday [1] and the parameters of the model are estimated using the classic Least Squares method. An appropriate metric is introduced in order to measure the error between the observed and the predicted distributions. In particular, the Wasserstein distance is proposed. Some properties of such metric are exploited to predict the response variable as direct linear combination of other independent histogram variables. Measures of goodness of fit are discussed. An application on real data corroborates the proposed method.

Keywords: Modal Symbolic Variables, Probability distribution function, histogram data, Ordinary Least Squares, Wasserstein distance

1. Introduction

In this paper we present a linear regression model for modal symbolic data in the framework of Symbolic Data Analysis (SDA).

Since the first pioneering papers of Edwin Diday, SDA has become a new statistical field of research gathering contributions from different scientific communities: statistics, knowledge extraction, machine learning, data mining. Three European projects supported the systematic developments of the main methodological contributions and a wide collection of SDA methods is available in three reference books (Bock and Diday [1], Billard and Diday [2] and Diday and Noirhomme-Fraiture [3]). Moreover, many other papers have been published on Journals and Conference proceedings and currently SDA is almost always in the list of the topics of the most relevant international data analysis conferences. The basic ideas of SDA consist of analyzing data which are described by set-valued variables. Symbolic data refer to a description of a class, a category, a set of individuals and, more generally, they correspond to a description of a “concept”. Examples of symbolic data are: football teams (classes of individuals); species of animals (categories); towns (concepts). Differently from the classic data where each “punctual” observation assumes only one value for every variable, symbolic data take multiple values, such as

Email addresses: antonio.irpino@unina2.it (Antonio Irpino), rosanna.verde@unina2.it (Rosanna Verde)

Preprint submitted to To be defined

February 8, 2012

intervals of continuous variables, different categories of nominal variables, empirical distribution or probability functions.

Symbolic data are receiving more and more attention because they are able to summarize huge sets of data, nowadays available in large databases or generated in streaming by smart meters or sensors.

A peculiarity of Symbolic data is that they allow to keep the variability in the data description, considering the interval of values that each observation can assume or the distribution of values. In this way, the methodological approaches developed in this context of analysis must guarantee this information is preserved.

Symbolic Data Analysis methods generalize multivariate data analysis to that new kind of data and they can be classified at least in three categories, according to the *input data* - the *method* - the *output data*, as follows:

symbolic-numerical-numerical: symbolic data in input are transformed into standard data in order to apply classic multivariate techniques. The results are classic data. For example, a dissimilarity between interval data is computed considering only the bounding values (minimum and maximum) of the intervals. That is a standard dissimilarity between punctual data and the measure is a single value.

symbolic-numerical-symbolic: symbolic data in input are analyzed according to a classic multivariate techniques and the results are symbolic data. For example, intervals are transformed in mid-points and radii, a classic analysis (e.g. linear regression) is performed on these data but the results (e.g. predictive variable) are furnished in terms of intervals by reconstructing them from the estimated mid-points and radii.

symbolic-symbolic-symbolic: symbolic data in input are transformed using generalization/specialization operators and the results are symbolic data. For example, symbolic data represented by intervals, histograms or distributions are aggregated in homogeneous classes through a clustering methods using a criterion of homogeneity which takes into account the characteristics of the data (internal structure). The results are classes of symbolic data expressed by the same kind of variables of the symbolic input data.

Most part of the SDA techniques are *symbolic-numerical-symbolic*, where the input and output data are symbolic and the methods are generalization of the classic data analysis methods to this new kind of data. Considerable contributions have been given in retrieving variability information of the symbolic data through graphic representations of the results and tools for the interpretation. A large overview of the SDA methods is in Bock and Diday [1] and in Diday and Noirhomme-Fraiture [3]. The linear regression models proposed in this context of analysis have been introduced to study the structure of dependence of a response *symbolic* variable from a set of independent or explicative variables of the same nature.

The first proposals were regression models for interval data as extension of the linear dependence model for numerical variables. Billard and Diday (2000) proposed a non-probabilistic approach based on the minimization of a criterion like the sum of squared error for the parameters estimation. The method consists in a regression model on the mid-points of the intervals and it makes reference to the basic statistics (mean, variance, correlation) introduced by Bertrand and Goupil [4] for interval data. de A. Lima Neto et al. [5], de A. Lima Neto and de A. T. de Carvalho [6] and de A. Lima Neto and de A. T. de Carvalho [7] improved the performance of the linear method adding the ranges of the intervals to the mid-points information in order to include the

variability expressed by the interval data. Therefore, the authors show the differences between this model and two regression models on the bounds of the intervals. Lima Neto et al. (2005) introduced an order constrained on the bounds of the predicted intervals in the model estimation process in order to guarantee the coherence between the observed and the predicted response variable. To overcome the problem of possible inversion of the bounds of the predicted intervals Lima Neto and De Carvalho (2008) suggested a non linear regression model on the mid-points and the ranges of the interval. Billard and Diday [2] presented at first a regression model for histogram variables. This approach is based on the basic statistics: mean, variance and correlation defined by Bertrand and Goupil [4] for intervals when they are assumed as random variables uniformly distributed. According to this approach the fitted regression model and their predicted values are generally single valued. The authors leave as an open problem output predicted values as symbolic data.

Verde and Irpino [8] proposed a simple linear regression model which allows to estimate a histogram response variable as linear transformation of another independent histogram variable. The main idea is to propose a suitable metric to measure the sum of squared errors between the observed and predicted multi-valued data (histograms or distributions). The Wasserstein distance [9] ℓ is a distance function defined between the probability distributions of two random variables X and Y , on a given metric space M . The minimal L_1 -metric ℓ_1 had been introduced and investigated already in 1940 by Kantorovich for compact metric spaces Kantorovich [10]. In 1914 Gini introduced the ℓ metric in a discrete setting on the real line [11] and Salvemini 1943 (in the discrete case) [12] and Dall'Aglio 1956 (in the general case) [13] proved the basic representation of L_p norm ℓ_p between the quantile functions of the two random variables. Mallows [14] introduced the ℓ_2 -metric in a statistical context. Moreover, starting from Mallows' work Bickel and Freedman [15] described topological properties and investigated applications to statistical problems as the bootstrap. They introduced the notion Mallows metric for ℓ_2 . So the L_p -metric ℓ_p was invented historically several times from different perspectives. Historically the notion of Gini – Dall' Aglio – Kantorovich – Wasserstein – Mallows metric would be correct for this class of metrics.

This measure that, we refer to henceforth as Wasserstein metric and that was already proposed by the authors in Clustering methods for interval [16] and histogram data [17, 18], seems particularly adapt in this context. According to this distance function, we study the dependence relationship of the histogram response variable from the explicative one considering the respective quantile functions.

Dias and Brito [19] referring to this last approach proposed a linear regression model for histogram data, directly interpreting the linear relationship between quantile functions. In the multiple regression model, as we will show in the present paper, one of the main problem is OLS cannot guarantee all the estimated parameters are positive. It is worth nothing that the predicted response variable is again a quantile function only if it is a linear combination of quantile functions with positive coefficients. In order to overcome such inconvenience Dias and Brito [19] proposed to introduce the so called symmetric quantile distributions in the model as new predictor variables. However the meaning of these new variables is not immediately interpretable. In the same paper a new measure of goodness-of-fit associated to the proposed model is also introduced.

Differently, our proposal is to exploit the properties of a decomposition of the Wasserstein proposed distance by Irpino and Romano [20], that is used to measure the sum of squared errors and rewrite the model splitting the contribution of the predictors in a part depending from the averages of the distributions and another depending from the centered quantile distributions.

The parameters associated to the predictors, constituted by the averages of the distributions, can be indifferently positive or negative because they effect only on the shift of the predicted quantile distribution. The authors already demonstrated for the simple regression model Verde and Irpino [8] that this leads to guarantee the positiveness of the parameter associated to the centered quantile function of the only predictor. However, in the multiple regression model, this solution is not automatically obtained, so that it needs to force the the positiveness of the multi parameters estimation by a Non Negative Least Squared algorithm. That is a classic algorithm that finds the solution among all the subsets of suboptimal OLS solutions. The rest of paper is organized as follows: in section 2 symbolic data are presented according to the definition given in Bock and Diday [1] and Diday and Noirhomme-Fraiture [3] books; in section 3 regression models for Numerical Probabilistic (Modal) Symbolic Variables are introduced and details on the new proposal are furnished; in section 4 some goodness of fit indices are proposed; in section 5 applications on real data are presented in order to corroborate the procedure.

2. Numerical symbolic data

Symbolic data allow to describe concepts, individuals or classes of individuals, by means of multiple values for each descriptor (variable). The term *symbolic variable* was coined in order to introduce such new set-valued descriptions. In a classic data table ($n \times p$ individuals per variables) each individual is described by a vector of values, similarly, in a *symbolic data table* each individual is described by a vector of set-valued descriptions (like intervals of values, histograms, set of numbers or of categories, sometimes equipped with weights, probabilities, frequencies, an so on). According to the taxonomy of symbolic variables presented in Bock and Diday [1] and recalled by Noirhomme-Fraiture and Brito [21], we may consider as numerical symbolic variables all those symbolic variables whose support is numeric.

Given a set of n individuals (concepts, classes) $\Omega = \{\omega_1, \dots, \omega_n\}$ a *symbolic variable* X with domain D is a map

$$X : \Omega \rightarrow D \quad X(\omega_i) \in D.$$

The different kinds of variable definitions depend on the nature of D . Considering only numerical domains, we can define the following symbolic variable:

Classic Variable It is observed when $D \subseteq \mathfrak{R}$, i.e. each individual ω_i is described by a single numeric value for the variable X ;

Interval Variable It is observed when $D \subseteq I\mathbb{R}$, where $I\mathbb{R}$ is the set of all intervals of real numbers $[a, b]$ where $a, b \in \mathfrak{R}$ and $a \leq b$;

Modal Variables According to Bock and Diday [1] the domains of *Modal Variables* are sets of mappings. Considering different kinds of mapping, several kinds of *Modal Symbolic Variables* can be defined. let us consider $D \subseteq M$ where $M_i \in M$ is a map $M_i : S_i \rightarrow W_i$, such that for each element of the support $s_i \in S_i$ it is associated $w_i = M_i(s_i) \in \mathfrak{R}^+$. If $M_i(s_i)$ has the same properties of a random variable (i.e. $\int_{s \in S_i} w(s) ds = 1$, or $\sum_{s \in S_i} w_s = 1$), X can be defined as a *Numerical Probabilistic (Modal) Symbolic Variable (NPSV)* and $M_i(s_i)$ can be described through a probability density function $f_i(x)$. Particular cases of such data arise when the generic individual ω_i is described by a model of random variables, a histogram, an empirical frequency distribution. In this paper we refer only to such kinds of data that we call *Numerical Probabilistic Symbolic Data (NPSD)*, that are in the domain of *Numerical Probabilistic (Modal) Symbolic Variables*.

For example, if ω_i for variable X is described by a normal distribution with parameters μ_i and σ_i , we may describe it by its probability density function (*pdf*) $f_i(x)$ as follows:

$$X(\omega_i) = f_i(x) = \{N(\mu_i, \sigma_i)\}.$$

Using the same notation, and according to Bertrand and Goupil [4] and Billard and Diday [2], we may consider interval data as a particular case of NPSD, where the *pdf* is uniform. Given an interval description of ω_i as $X(\omega_i) = [a_i, b_i]$, we may rewrite the same description in terms of NPSD as:

$$X(\omega_i) = f_i(x) = \{U(a_i, b_i)\}.$$

Histogram data are a particular case of NPSD, where, given the generic individual ω_i , a set of disjoint K_i intervals $I_{ki} = [a_{ki}, b_{ki}]$ $k = 1, \dots, K_i$ and a set of positive K_i weights w_{ki} such that $\sum_{k=1}^{K_i} w_{ki} = 1$, its description for the NPSV X is :

$$X(\omega_i) = f_i(x) = \{(I_{1i}, w_{1i}), \dots, (I_{ki}, w_{ki}), \dots, (I_{K_i i}, w_{K_i i})\}.$$

Also in this case it is possible to define a *pdf* for each histogram data as proposed by Irpino and Verde [17], considering a histogram as a mixture of uniform *pdf*'s.

In another way an interval can be treated as a histogram with $k = 1$, such that $I_{1i} = [a_i, b_i]$ and $w_{1i} = 1$.

Similarly, classic data (single valued numerical data), can be considered as NPSD, where the description is a *Dirac delta function* or as histogram data with one thin ($a_i = b_i$) interval.

Notation and definitions. Let $X_1, \dots, X_j, \dots, X_p$ and Y be the independent and dependent NPSV's observed on a set Ω of n individuals (concepts or classes). We denote with:

- $f_i(x_j)$ and $f_i(y)$ the empirical or theoretical *probability density functions* (*pdf*'s), i.e. NPSD describing the i -th individual (for $i = 1, \dots, n$);
- $F_i(x_j)$ and $F_i(y)$ the *cumulative distribution functions* (*cdf*'s);
- $x_{ij}(t) = F_i^{-1}(x_j)$ and $y_i(t) = F_i^{-1}(y)$ the *quantile functions* (*qf*'s);
- \bar{x}_{ij} , \bar{y}_i and s_{ij} , s_i^y , the *means* and the *standard deviations* of the $x_{ij}(t)$'s and of $y_i(t)$ respectively. They are real numbers;
- $\bar{x}_j(t) = \frac{1}{n} \sum_{i=1}^n x_{ij}(t)$, $\bar{y}(t) = \frac{1}{n} \sum_{i=1}^n y_i(t)$ the means of the sets of n distributions $x_{ij}(t)$ and $y_i(t)$, i.e. the *baricenter distributions*;
- $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n \int_0^1 x_{ij}(t) dt = \frac{1}{n} \int_0^1 \bar{x}_j(t) dt$, $\bar{y} = \sum_{i=1}^n \int_0^1 \bar{y}(t) dt = \bar{y} = \int_0^1 \bar{y}(t) dt$ the means of the distribution means of the $x_{ij}(t)$'s and $y_i(t)$, or equivalently the means of the baricenter distributions of the $x_{ij}(t)$'s and $y_i(t)$. They are real numbers;
- $x_{ij}^c(t) = x_{ij}(t) - \bar{x}_{ij}$ the *centred quantile function*, i.e. the quantile function shifted by $-\bar{x}_{ij}$.

$$^1 \bar{x}_{ij} = \int_0^1 x_{ij}(t) dt, \bar{y}_i = \int_0^1 y_i(t) dt \text{ and } s_{ij} = \sqrt{\int_0^1 x_{ij}^2(t) dt - \bar{x}_{ij}^2}, s_i^y = \sqrt{\int_0^1 y_i^2(t) dt - \bar{y}_i^2}$$

3. OLS linear regression for NPSD

Given X_1, \dots, X_p p explicative NPSV's and a dependent NPSV Y observed on a set Ω , the aim is to fit the parameters of a linear regression function $\phi(X)$. Denoted with \mathbf{X} and \mathbf{Y} respectively the matrix collecting the observed values of the X_j explicative NPSV's and the vector of the observed values of the predictive NPSV Y , we write the regression model as follows:

$$\mathbf{Y} = \phi(\mathbf{X}) + \varepsilon \quad (1)$$

As for classic data, and according to the definitions of NPSD's, the model is fitted starting from the following symbolic data table, where instead of a matrix of scalar values, we have a matrix of NPSD's:

$$[\mathbf{Y}|\mathbf{X}] = \left[\begin{array}{c|cccc} f_1(y) & f_1(x_1) & \cdots & f_1(x_j) & \cdots & f_1(x_p) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ f_i(y) & f_i(x_1) & \cdots & f_i(x_j) & \cdots & f_i(x_p) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ f_n(y) & f_n(x_1) & \cdots & f_n(x_j) & \cdots & f_n(x_p) \end{array} \right]. \quad (2)$$

Two main approaches for the estimation of the parameters of linear regression model have been proposed when the symbolic data are histograms. Starting from the elementary statistics proposed by Bertrand and Goupil [4], in a first approach Billard and Diday [2] introduced an extension of the classic OLS (Ordinary Least Squares) linear regression model to the histogram-valued variables. A second group of approaches is based on the use of the quantile functions (which is biunivocal to a *pdf*) of the NPSD and of the Wasserstein distance for defining the sum of square errors in the OLS problem. The idea behind the latest approaches is to predict a quantile function after having observed a set of quantile functions as predictors.

3.1. The Billard-Diday model

According to Billard and Diday [2], the regression model which expresses a linear relationship between a set of predictors and a response histogram variable is based on the assumptions of Bertrand and Goupil [4], they consider a histogram as the representation of a cluster of individuals. A second implicit assumption is that the histograms are the marginal distributions of a multivariate distribution with independent components: i.e., given the i -th description, and the two $f_i(x)$ and $f_i(y)$ *pdf*'s, and the joint *pdf* is expressed as $f_i(x, y) = f_i(x) \cdot f_i(y)$. In this approach, interval data are considered as uniform *pdf*'s, and as a particular case of histogram with just one interval with unitary weight. The regression method is based on the identification of the covariance matrix that depends on the following basic statistics; being Y and X_j the NPSV observed for n individuals, the means and the variances of each variable are computed according to:

$$\bar{y} = \int_{-\infty}^{+\infty} y \cdot f(y) dy = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{+\infty} y \cdot f_i(y) dy = \frac{1}{n} \sum_{i=1}^n \bar{y}_i \quad (3)$$

$$s_y^2 = \int_{-\infty}^{+\infty} y^2 \cdot f(y) dy - \bar{y}^2 = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{+\infty} y^2 \cdot f_i(y) dy - \bar{y}^2 \quad (4)$$

and considering that

$$f(x, y) = \frac{1}{n} \sum_{i=1}^n f_i(x, y) = \frac{1}{n} \sum_{i=1}^n f_i(x) \cdot f_i(y) \quad (5)$$

the covariance measure proposed by Bertrand and Goupil [4] is

$$s_{x,y} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x \cdot y \cdot f(x, y) dx dy - \bar{x} \cdot \bar{y} = \frac{1}{n} \sum_{i=1}^n \bar{x}_i \bar{y}_i - \bar{x} \cdot \bar{y}. \quad (6)$$

It is important to note that in general $s_x^2 \neq s_{x,x}$. Billard and Diday [2] proposed a different way to compute $s_{x,y}$ when data are intervals, considering them as uniform distributions as well as when data are histograms, considering them as weighted intervals.

The proposal extends the linear regression model for standard data, according to the following equation:

$$y = \beta_0 + \sum_{j=1}^p \beta_j x_j + e \quad (7)$$

where β' s are estimated by means of the classic OLS estimators as follows:

$$\begin{bmatrix} \hat{\beta}_1 \\ \dots \\ \hat{\beta}_j \\ \dots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} s_{x_1}^2 & \dots & s_{x_1, x_j} & \dots & s_{x_1, x_p} \\ \dots & \dots & \dots & \dots & \dots \\ s_{x_j, x_1} & \dots & s_{x_j}^2 & \dots & s_{x_j, x_p} \\ \dots & \dots & \dots & \dots & \dots \\ s_{x_p, x_1} & \dots & s_{x_p, x_j} & \dots & s_{x_p}^2 \end{bmatrix}^{-1} \begin{bmatrix} s_{y, x_1} \\ \dots \\ s_{y, x_j} \\ \dots \\ s_{y, x_p} \end{bmatrix}$$

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^p \hat{\beta}_j \bar{x}_j$$

The estimated model allows to predict \hat{y}_i as follows

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{ij} \quad (8)$$

Therefore, that is a method for predicting single values but not directly distributions (this point is also considered by the authors). In general, it is difficult to express the distribution of a linear combination of random variables, in particular when the random variables are not identically distributed. In general, an approximation of the distribution associated to \hat{y} can be obtained by means of a Montecarlo experiment.

On the other hand, if a set of strong conditions hold (the knowledge of the cardinality of groups, the internal independence of the multivariate distribution in each group), the model parameters have the same inferential properties of the classic OLS linear regression estimation method.

3.2. Wasserstein distance based models

We have recalled that the Billard and Diday [2] model is implicitly founded on the modeling of the union of groups or concepts and it is mainly based on the basic statistics proposed by Bertrand and Goupil [4]. For example if there are two groups G_1 and G_2 of people with the same cardinality, described by their income distributions $f_1(INCOME)$ and $f_2(INCOME)$, all the basic statistics correspond to the classic statistics calculated for the mixture of distributions. From this point of view the basic statistics of the group $G_1 \cup G_2$ are those calculated considering the *pdf* of the variable *INCOME* as: $f(INCOME) = 0.5f_1(INCOME) + 0.5f_2(INCOME)$. This situation arises frequently when the aim is to describe unions of groups (for example, municipalities are grouped into cities). In other cases, this approach can be inconsistent. For example, we cannot know the cardinality of the groups or it makes no sense to know the number of the elementary observations: if we take several pulse rate measurements of two individuals ω_1 and ω_2 and we fit a distribution or a histogram $f_1(PulseRate)$ and $f_2(PulseRate)$ for each one of them, we may be interested to discover relations between the two individuals by means of the comparison of their (probabilistic) respective distributions, instead of considering a mixture of distributions (that is also a logical non sense, two individuals cannot be fused into a super individual!). In this sense Verde and Irpino [18] proposed a different approach based on the comparison of distributions by means of suitable dissimilarity measures. Verde and Irpino [18] considered different kinds of probabilistic metrics for histogram data and suggested that the same results can be extended to data described by density functions (i.e., NPSD). Among the discussed metrics, the ℓ_2 Wasserstein [9] distance permits to explain and interpret in an easy way the proximity relations between two probability functions. Given two *pdf*'s $f(x)$ and $g(x)$, with means \bar{x}_f and \bar{x}_g , finite standard deviations s_f and s_g it is possible to associate respectively their *cdf*'s $F(x)$ and $G(x)$. With each *cdf*'s it is associated their *quantile functions* (*qf*), i.e. the inverse functions of the *cdf*: $x_f(t) = F^{-1}(t)$ and $x_g(t) = G^{-1}(t)$. The ℓ_2 Wasserstein distance is the following:

$$d_W(f, g) = \sqrt{\int_0^1 [x_f(t) - x_g(t)]^2 dt}. \quad (9)$$

The ℓ_2 Wasserstein distance is proposed for calculating the square errors in the OLS problem. Given the matrix (2), we consider the associated matrix \mathbf{M} containing the corresponding quantile functions:

$$\mathbf{M} = \left[\begin{array}{c|cccc} \mathbf{Y} & \mathbf{X} \end{array} \right] = \left[\begin{array}{c|cccc} y_1(t) & x_{11}(t) & \cdots & x_{1j}(t) & \cdots & x_{1p}(t) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_i(t) & x_{i1}(t) & \cdots & x_{ij}(t) & \cdots & x_{ip}(t) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_n(t) & x_{n1}(t) & \cdots & x_{nj}(t) & \cdots & x_{np}(t) \end{array} \right]; \quad (10)$$

In this case, given a set of p quantile functions for the i -th individual, we look for a linear combination of $x_{ij}(t)$'s (for $j = 1, \dots, p$) which allows to predict the $y_i(t)$'s (for $i = 1, \dots, n$) except for an error term $e_i(t)$. It is worth noting that $e_i(t)$ is a residual function, not necessarily a quantile function. The model to be fit is the following:

$$y_i(t) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}(t) + e_i(t), \quad (11)$$

where the Sum of Squared Errors (SSE) to minimize for the solution of the OLS problem is related to the Squared ℓ_2 Wasserstein distance as follows:

$$SSE = \sum_{i=1}^n [e_i(t)]^2 = \sum_{i=1}^n d_W^2 \left(y_i(t), \left[\beta_0 + \sum_{j=1}^p \beta_j x_{ij}(t) \right] \right). \quad (12)$$

A problem arises for the linear combination of quantile functions: only if $\beta_j \geq 0$ ($j = 1, \dots, p$) it is assured that $y_i(t)$ is a quantile function (i.e. a not decreasing function). In order, to overcome this problem, Dias and Brito [19] proposed a novel method for the regression of histogram valued data based on the Wasserstein distance between quantile functions. In order to take into account also inverse casual relations, Dias and Brito [19] proposed to expand the matrix \mathbf{M} adding also the quantile functions of the symmetric distributions of the explicative symbolic variables. Given $f_i(x_j)$ (with the respective quantile function $x_{ij}(t)$), the corresponding symmetric distribution $\tilde{f}_i(x_j)$ (and its quantile function $\tilde{x}_{ij}(t)$) is obtained by multiplying the support of $f_i(x_j)$ by -1 , such that the integral of the sum of the two quantile functions is equal to zero ($\int_0^1 [x_{ij}(t) + \tilde{x}_{ij}(t)] dt = 0$).

The model is the following:

$$y_i(t) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}(t) + \sum_{j=1}^p \tilde{\beta}_j \tilde{x}_{ij}(t) + e_i(t), \quad (13)$$

and the estimation of the parameters is obtained optimizing the following constrained OLS problem:

$$\begin{aligned} \underset{(\beta_0, \beta_j, \tilde{\beta}_j)}{\operatorname{argmin}} SS &= \sum_{i=1}^n d_W^2 \left(y_i(t), \left[\beta_0 + \sum_{j=1}^p \beta_j x_{ij}(t) + \sum_{j=1}^p \tilde{\beta}_j \tilde{x}_{ij}(t) \right] \right) \\ \text{s.t.} \quad &\beta_j, \tilde{\beta}_j \geq 0 \end{aligned}$$

3.3. The Irpino and Verde model for multiple regression

The negative value of the parameter β_j in model (11) is in general not acceptable when dealing with quantile functions. In order to overcome this inconvenient, starting from a particular decomposition of the Wasserstein Verde and Irpino [8] presented a new formulation of the problem in the simple regression model, that is when only one variable X affects the predicted variable Y . An extension of this approach to the multivariate case has been already presented by the authors [22] and it takes into consideration the matrices in equation (19).

The introduction of the new model can be done according to some preliminary considerations about the properties of ℓ_2 Wasserstein distance decomposition. Given f and g two NPSD and $x_f^c(t)$ and $x_g^c(t)$ the respective centred quantiles functions (above defined in §2 *Notation and definitions*), Cuesta-Albertos et al. [23] showed that the ℓ_2 Wasserstein distance can be rewritten as

$$d_W^2(f, g) = (\bar{x}_f - \bar{x}_g)^2 + \int_0^1 [x_f^c(t) - x_g^c(t)]^2 dt. \quad (14)$$

This property allows to consider the squared distance as the sum of two components, the first related to the location of NPSD and the second related to their variability structure. Irpino and

Romano [20] improved such decomposition, showing that the d_W^2 can be finally decomposed into three quantities:

$$d_W^2(f, g) = (\bar{x}_f - \bar{x}_g)^2 + (s_f - s_g)^2 + 2s_f s_g (1 - \rho(x_f, x_g)) \quad (15)$$

where $\rho(x_f, x_g)$ is a correlation coefficient about the quantile functions, i.e.:

$$\rho(x_f, x_g) = \frac{\int_0^1 x_f(t) \cdot x_g(t) dt - \bar{x}_f \cdot \bar{x}_g}{s_f \cdot s_g}. \quad (16)$$

Irpino et al. [24] showed some computational aspects relating to ρ when data are histograms and they showed that ρ is computed in a linear time with the total number of bins of the histograms. Moreover, the Equation 16 allows to define the inner product between two qf 's, as follows:

$$\langle x_f(t), x_g(t) \rangle = \int_0^1 x_f(t) \cdot x_g(t) dt = \rho(x_f, x_g) \cdot s_f \cdot s_g + \bar{x}_f \cdot \bar{x}_g. \quad (17)$$

Thus, given two vectors of quantile functions $\mathbf{x} = [x_i(t)]_{n \times 1}$ and $\mathbf{y} = [y_i(t)]_{n \times 1}$, we can define the scalar product of two vectors of NPSD as:

$$\mathbf{x}^T \mathbf{y} = \sum_{i=1}^n [\rho(x_i, y_i) \cdot s_{x_i} \cdot s_{y_i} + \bar{x}_i \cdot \bar{y}_i]. \quad (18)$$

If we consider that $x_{ij}^c(t) = x_{ij}(t) - \bar{x}_{ij}$, each element of \mathbf{X} can be rewritten as $x_{ij}(t) = x_{ij}^c(t) + \bar{x}_{ij}$. The same is valid for vector \mathbf{Y} . Matrix \mathbf{M} is transformed as:

$$\mathbf{M} = \begin{bmatrix} \bar{\mathbf{Y}} + \mathbf{Y}^c & | & \bar{\mathbf{X}} + \mathbf{X}^c \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{Y}} & | & \bar{\mathbf{X}} \end{bmatrix} + \begin{bmatrix} \mathbf{Y}^c & | & \mathbf{X}^c \end{bmatrix} \quad (19)$$

where $\bar{\mathbf{Y}} = [\bar{y}_i]_{n \times 1}$ is the vector of the means of the $f_i(y)$, $\mathbf{Y}^c = [y_i^c(t)]_{n \times 1}$ is the vector of the centred quantile functions of $f_i(y)$'s, $\bar{\mathbf{X}} = [\bar{x}_{ij}]_{n \times p}$ is the matrix of the means of the $f_i(x_j)$, $\mathbf{X}^c = [x_{ij}^c(t)]_{n \times p}$ is the matrix of the centred quantile functions of $f_i(x_j)$'s.

We assume that each quantile function $y_i(t)$ can be expressed as a linear combination of the means \bar{x}_{ij} and of the centred quantile functions $x_{ij}^c(t)$ plus an error term $e_i(t)$ (which is a function) as follows:

$$y_i(t) = \beta_0 + \sum_{j=1}^p \beta_j \bar{x}_{ij} + \sum_{j=1}^p \gamma_j x_{ij}^c(t) + e_i(t) \quad (20)$$

If we consider the matrix $\bar{\mathbf{X}}_+ = [\mathbf{1} | \bar{\mathbf{X}}]$, we can rewrite the model in (20) using the matrix notation as follows:

$$\mathbf{Y} = \bar{\mathbf{X}}_+ \mathbf{B} + \mathbf{X} \mathbf{\Gamma} + \mathbf{e} \quad (21)$$

In order to estimate the parameters, we define the Sum of Square Errors function (SSE) like in the OLS method, using the Wasserstein ℓ_2 measure:

$$SSE = \sum_{i=1}^n \int_0^1 e_i^2(t) dt = \sum_{i=1}^n \int_0^1 \left[y_i(t) - \beta_0 + \sum_{j=1}^p \beta_j \bar{x}_{ij} + \sum_{j=1}^p \gamma_j x_{ij}^c(t) \right]^2 dt \quad (22)$$

in matrix form:

$$SSE(\mathbf{B}, \mathbf{\Gamma}) = \mathbf{e}^T \mathbf{e} = [\mathbf{Y} - \bar{\mathbf{X}}_+ \mathbf{B} - \mathbf{X}^c \mathbf{\Gamma}]^T [\mathbf{Y} - \bar{\mathbf{X}}_+ \mathbf{B} - \mathbf{X}^c \mathbf{\Gamma}] \quad (23)$$

Considering the equation (21) induced from the ℓ_2 Wasserstein metric, we have $\bar{\mathbf{X}}_+^T \mathbf{X}^c = \mathbf{0}_{(p+1) \times p}$, $\bar{\mathbf{X}}_+^T \mathbf{Y} = \bar{\mathbf{X}}_+^T \bar{\mathbf{Y}}$ and $\mathbf{X}^{cT} \mathbf{Y} = \mathbf{X}^{cT} \mathbf{Y}^c$. Then, SSE in equation (23) can be decomposed into two positive quantities as follows:

$$SSE(\mathbf{B}, \mathbf{\Gamma}) = SSE(\mathbf{B}) + SSE(\mathbf{\Gamma}) = \bar{\mathbf{e}}^T \bar{\mathbf{e}} + (\mathbf{e}^c)^T \mathbf{e}^c \quad (24)$$

where:

$$\bar{\mathbf{e}} = \bar{\mathbf{Y}} - \bar{\mathbf{X}}_+ \mathbf{B} \quad (25)$$

$$\mathbf{e}^c = \mathbf{Y}^c - \mathbf{X}^c \mathbf{\Gamma} \quad (26)$$

with $\bar{\mathbf{e}} = [\bar{\mathbf{e}}_i]_{n \times 1}$ a vector of real numbers.

We may express the single minimization problem as the minimization of two independent functions: the first one related to the means of the predictor quantile functions \bar{x}_{ij} 's in $\bar{\mathbf{X}}_+$, and the second one related to the variability of the centered quantile distributions $x_{ij}^c(t)$'s in \mathbf{X}^c . Then two models are independently estimated:

$$\bar{\mathbf{Y}} = \bar{\mathbf{X}}_+ \mathbf{B} + \bar{\mathbf{e}} \quad (27)$$

$$\mathbf{Y}^c = \mathbf{X}^c \mathbf{\Gamma} + \mathbf{e}^c \quad (28)$$

The first can be solved as classic OLS problem for the estimation of \mathbf{B} :

$$\underset{\mathbf{B}}{\operatorname{argmin}} SSE(\mathbf{B}) = [\bar{\mathbf{Y}} - \bar{\mathbf{X}}_+ \mathbf{B} - \bar{\mathbf{e}}]^T [\bar{\mathbf{Y}} - \bar{\mathbf{X}}_+ \mathbf{B} - \bar{\mathbf{e}}] \quad (29)$$

The second (30) is solved using the NNLS (Non Negative Least Squares) algorithm proposed by Lawson and Hanson [25], modified with the introduction of the product between quantile functions (Eq. 17) in the classic matrix computations:

$$\underset{\mathbf{\Gamma}}{\operatorname{argmin}} SSE(\mathbf{\Gamma}) = [\mathbf{Y}^c - \mathbf{X}^c \mathbf{\Gamma} - \mathbf{e}^c]^T [\mathbf{Y}^c - \mathbf{X}^c \mathbf{\Gamma} - \mathbf{e}^c] \quad (30)$$

s.a. $\gamma_j \geq 0 \quad j = 1, \dots, p.$

The estimated OLS parameters are:

$$\hat{\mathbf{B}} = (\bar{\mathbf{X}}_+^T \bar{\mathbf{X}}_+)^{-1} \bar{\mathbf{X}}_+^T \bar{\mathbf{Y}} \quad (31)$$

$$\hat{\mathbf{\Gamma}} = (\mathbf{X}^{cT} \mathbf{X}^c)^{-1} \mathbf{X}^{cT} \mathbf{Y}^c. \quad (32)$$

Therefore, a *qf* $\hat{y}_i(t)$ is predicted by the estimated linear model according to the estimated parameters:

$$\hat{y}_i(t) = \hat{\beta}_0 + \hat{y}_i^c(t) = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j \bar{x}_{ij} + \sum_{j=1}^p \hat{\gamma}_j x_{ij}^c(t). \quad (33)$$

²Some algebraic details are in appendix A

4. The goodness of fit evaluation

Considering the nature of the data, the evaluation of the goodness of fit of the model is not straightforward like for the classic linear regression models. Here we present three indices that can be used for evaluating the goodness of fit of a regression model on NPSD. The first is the Ω measure proposed by Dias and Brito [19], the second is the *Pseudo* – R^2 proposed by Verde and Irpino [8] and the last is the classic square root of the mean sum of squares expressed using the ℓ_2 Wasserstein distance (we denote it as $RMS E_W$). The three measures are detailed in the following.

Ω Dias and Brito [19] The proposed measure is the ratio

$$\Omega = \frac{\sum_{i=1}^n d_W^2(\hat{y}_i(t), \bar{y})}{\sum_{i=1}^n d_W^2(y_i(t), \bar{y})} \quad (34)$$

that varies from zero to 1.

Pseudo – R^2 It is a measure proposed by Verde and Irpino [8] for the simple linear regression model. Verde and Irpino [26] proved that the Wasserstein distance can be used for the definition of a sum of squared deviation as follows:

$$SSY = n \cdot s_y^2 = \sum_{i=1}^n d_W^2(y_i(t), \bar{y}(t)) = \sum_{i=1}^n \int_0^1 [y_i(t) - \bar{y}(t)]^2 dt. \quad (35)$$

where $\bar{y}(t)$ is the *baricenter* of NPSD $y_i(t)$'s as defined at the end of section 2.

A common tool for the evaluation the goodness of fit of the model is the well known *coefficient of determination* R^2 ($R^2 = \frac{SSR}{SSY}$ or $R^2 = 1 - \frac{SSE}{SSY}$

Computation of this coefficient is based on the partitions of the total variation in the dependent variable, denoted SST, into two parts: the part explained by the estimated regression equation, denoted SSR, and the part that measures the unexplained variation, SSE, referred to as the residual sum of squares:

$$SSY = SSE + SSR \quad (36)$$

In our case, in general, the equality does not hold, and we prove³ that the decomposition of SSY is the following:

$$\begin{aligned} SSY &= \underbrace{\sum_{i=1}^n \int_0^1 [\hat{y}_i(t) - y_i(t)]^2 dt}_{SSE} + \underbrace{\sum_{i=1}^n \int_0^1 [\bar{y}(t) - \hat{y}_i(t)]^2 dt}_{SSR} \\ &\quad - 2 \underbrace{\left[n \cdot \left(\sigma_{\bar{y}}^2 - \sum_{j=1}^p \gamma_j r_{\bar{y}\bar{x}_j} \sigma_{\bar{y}} \sigma_{\bar{x}_j} \right) + \mathbf{\Gamma} \nabla S S(\mathbf{\Gamma}) \right]}_{Bias}. \end{aligned} \quad (37)$$

³See appendix Appendix B.

The *bias* term $\mathbf{I}\nabla\mathbf{SS}(\mathbf{I})$ in eq. (37) reflects the impossibility of the linear transformation of $\bar{x}(t)$ of reflecting the variability structure of $\bar{y}(t)$. In general, this term goes to zero when NPSD have the same shape (i.e., from the third ones forward, the standardized moments of the histograms are equal) and the standard deviations of $f_{ij}(x)$'s are proportional to the standard deviations of the $f_i(y)$'s. Further, the term $\mathbf{I}\nabla\mathbf{SS}(\mathbf{I})$ indicates the impact of the suboptimal solution obtained using the NNLS and is equal to zero only when the γ_j 's are the same of those calculated without the non negativeness constraint.

In this case, the classic $R^2 = 1 - \frac{SSE}{SSY}$ statistic can be less than zero or greater than 1. In order to obtain a measure of goodness of fit that does not suffer of the described drawback, we propose to adopt the following general index, that takes values in the real interval

$$0, 1$$

:

$$PseudoR^2 = \min \left[\max \left[0; 1 - \frac{SSE}{SSY} \right]; 1 \right]. \quad (38)$$

Denoting with *Bias* = $\mathbf{I}\nabla\mathbf{SS}(\mathbf{I})$, this index is also equal to:

$$PseudoR^2 = \frac{bias}{SSY}. \quad (39)$$

RMSE The Root Mean Square Error is generally used as measure of goodness of fit. Choosing an appropriate measure for computing the distance between NPSD's, we may compute the RMSE. In this paper, having used the Wassertein distance, we propose the following measure for the RMSE:

$$RMSE_W = \sqrt{\frac{\sum_{i=1}^n \int_0^1 (\hat{y}_i(t) - y_i(t))^2 dt}{n}} = \sqrt{\frac{SSE}{n}}. \quad (40)$$

5. Application on real data

To illustrate the proposed method we choose some examples presented in the literature on clinic data and a new climatic dataset. Expecially, we make use only of dataset of NPSD described by histograms usually arisen as summaries of large amount of data.

The first dataset is presented by Billard and Diday [2, Chap.6, Table 6.8] and also presented as application in Dias and Brito [19]. In the dataset presented in table 1 there are the Hematocrit (Y) histogram NPSD and the Hemoglobin (X) histogram NPSD observed for 10 units.

Fig. 1 shows the graphical representation of table 1 and the graphic representation of the means (barycenter) NPSD of each histogram variable, according to the barycenter of histogram variable as presented in [26].

Table 2 specifies the main summary statistics for the two histogram variables using the Billard and Diday [2] set of summary statistics and those proposed by Verde and Irpino [26]. For the barycenters the mean and the standard deviation are only reported. Obviously, it is possible to report also the other moments of the barycenters, but for the sake of brevity we prefer to leave to the reader further considerations looking directly to their graphical representations in fig. 1.

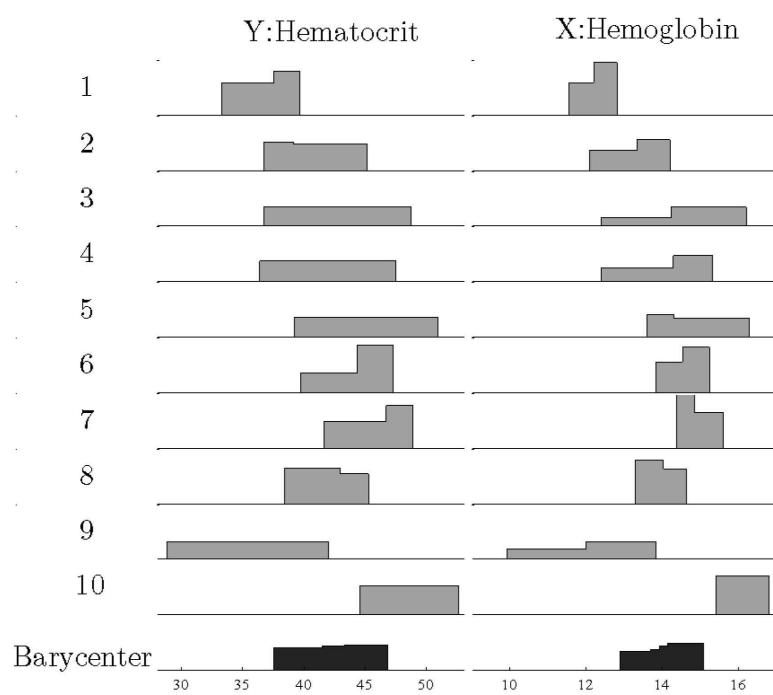


Figure 1: Blood dataset. Histogram representation, in the last row are represented the barycenter histograms.

Units	Y: Hematocrit	X: Hemoglobin
1	{{[33.29;37.57[, 0.6 ; [37.52; 39.61], 0.4 }	{{[11.54; 12.19[, 0.4 ; [12.19; 12.8], 0.6 }
2	{{[36.69; 39.11[, 0.3 ; [39.11; 45.12], 0.7 }	{{[12.07; 13.32[, 0.5 ; [13.32; 14.17], 0.5 }
3	{{[36.69; 42.64[, 0.5 ; [42.64; 48.68], 0.5 }	{{[12.38; 14.2[, 0.3 ; [14.2; 16.16], 0.7 }
4	{{[36.38; 40.87[, 0.4 ; [40.87; 47.41], 0.6 }	{{[12.38; 14.26[, 0.5 ; [14.26; 15.29], 0.5 }
5	{{[39.19; 50.86[, 1 }	{{[13.58; 14.28[, 0.3 ; [14.28; 16.24], 0.7 }
6	{{[39.7; 44.32[, 0.4 ; [44.32; 47.24], 0.6 }	{{[13.81; 14.5[, 0.4 ; [14.5; 15.2], 0.6 }
7	{{[41.56; 46.65[, 0.6 ; [46.65; 48.81], 0.4 }	{{[14.34; 14.81[, 0.5 ; [14.81; 15.55], 0.5 }
8	{{[38.4; 42.93[, 0.7 ; [42.93; 45.22], 0.3 }	{{[13.27; 14.0[, 0.6 ; [14.0; 14.6], 0.4 }
9	{{[28.83; 35.55[, 0.5 ; [35.55; 41.98], 0.5 }	{{[9.92; 11.98[, 0.4 ; [11.98; 13.8], 0.6 }
10	{{[44.48; 52.53[, 1 }	{{[15.37; 15.78[, 0.3 ; [15.78; 16.75], 0.7 }

Table 1: Blood dataset: 10 units described by two histogram-valued variables.

	Y:Hemoglobin	X: Hematocrit
n		10
Mean (BD)	42.26	14.05
Barycenter mean(VI)	42.26	14.05
Barycenter std (VI)	2.660	0.622
Standard deviation (BD)	4.658	1.355
Standard deviation (VI)	3.824	1.204
Correlation (BD)		0.903
Correlation (VI)		0.979

Table 2: Blood dataset: summary statistics. BD refers to the approach of [2], while VI refers to the approach of [26]

In Tabs. 3, 4 and 5, the rows labeled with *Observed* reports the OLS estimates of the parameters of the regression models using the formulation of Billard and Diday [2] as described in 3.1, the formulation of Dias and Brito [19] as presented in 3.2 and the novel formulation as proposed in 3.3.

In order to compare the three models, we computed the goodness of fit measures presented in 4. We remark that Billard-Diday model does not allow directly to predict a distribution function or a quantile function associated with $\hat{f}_i(y)$, thus, for computing the goodness of fit indices of the Billard-Diday model, we performed a Montecarlo experiment in order to estimate the predicted $\hat{f}_i(y)$ distribution.

In order to calculate the confidence interval of the parameters of the three models, and considering the complexity of establishing manageable probabilistic hypotheses on the error functions $e_i(t)$ like in the classic linear regression estimation problem, we have performed the bootstrap [27] estimates of the parameters of the models by constructing 1,000 resamples of the observed dataset (and of equal size to the observed dataset), each of which is obtained by random sampling with replacement from the original dataset. The confidence interval of each parameter is calculated using the percentile method, i.e. by using the 2.5 and the 97.5 percentiles of the bootstrap distribution as the limits of the 95% confidence interval for each parameter.

As usual, the point estimates is the mean bootstrap value. The main results about the parameters and the goodness of fit indices are shown in Tables 3,4 and 5.

The results show that the Dias-Brito and the novel proposed method fit better than the Billard-Diday model, and that the Dias-Brito and the proposed method have negligible differences con-

Billard-Diday Model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$					
	Model parameters			Goodness of fit	
	$\hat{\beta}_0$	$\hat{\beta}_1$	Ω	Pseudo R^2	$RMS E_W$
Observed	-1.336	3.103	0.811	0.919	3.431
Bootstrap estimates					
Mean	2.224	2.851	0.692	0.812	3.860
Bias	3.561	-0.253			
SE	6.592	0.461			
2.5%	-4.528	1.609	0.197	0.155	2.709
97.5%	20.243	3.318	0.888	0.953	6.854

Table 3: Blood dataset: Billard-Diday model parameters estimated on the full dataset and bootstrapping the dataset.

Dias-Brito Model: $\hat{y}_i(t) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i(t) + \hat{\beta}_1 \cdot \tilde{x}_i(t)$						
	Model parameters			Goodness of fit		
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\tilde{\beta}_1$	Ω	Pseudo R^2	$RMS E_W$
Observed	-1.953	3.560	0.413	0.963	0.945	3.431
Bootstrap estimates						
Mean	-1.657	3.574	0.448	0.963	0.935	2.623
Bias	0.296	0.014	0.035			
SE	2.862	0.165	0.164			
2.5%	-5.848	3.255	0.217	0.928	0.823	1.671
97.5%	5.037	3.931	0.848	0.986	0.981	3.345

Table 4: Blood dataset: Dias-Brito model parameters estimated on the full dataset and bootstrapping the dataset.

Irpino-Verde model: $\hat{y}_i(t) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \bar{x}_i + \hat{\gamma}_1 \cdot x_i^c(t)$						
	Model parameters			Goodness of fit		
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_1$	Ω	Pseudo R^2	$RMS E_W$
Observed	-2.157	3.161	3.918	0.961	0.943	2.892
Bootstrap estimates						
Mean	-1.928	3.146	3.969	0.961	0.931	2.688
Bias	0.229	-0.016	0.051			
SE	2.833	0.199	0.269			
2.5%	-6.348	2.688	3.602	0.924	0.812	1.730
97.5%	4.644	3.462	4.710	0.985	0.980	3.403

Table 5: Blood dataset: Irpino-Verde model parameters estimated on the full dataset and bootstrapping the dataset.

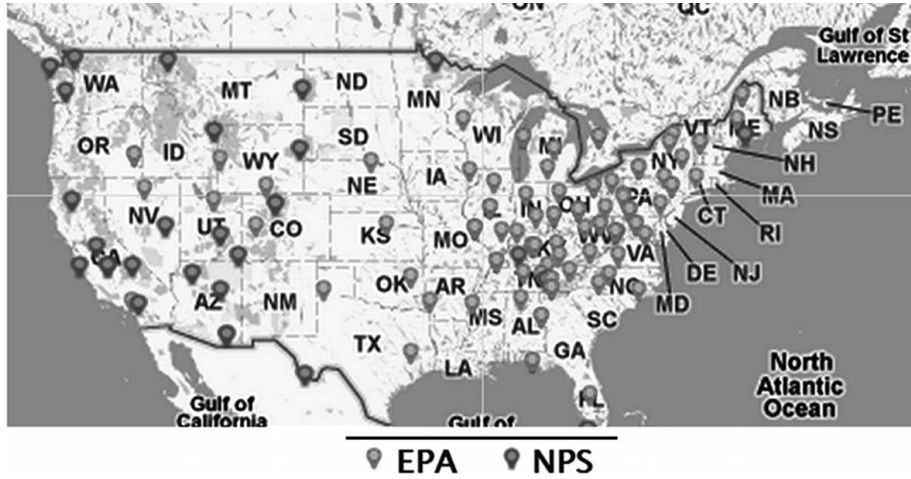


Figure 2: Ozone dataset. EPA and NPS monitored sites.

sidering the goodness of fit. The difference is about the interpretation of the regression parameters. The parameters of the Billard-Diday model are interpretable as for classic regression but the dependent predicted distributions cannot be easily described. The Dias-Brito model parameters show that a unitary (positive) variation of all the quantiles of the Hematocrit induces a variation of 3.574 of the Hemoglobin quantiles, while a unitary increase of the “symmetric” Hematocrit quantiles induces an increase of the Hemoglobin quantiles equal to 0.448, this can be a bit confusing considering that there is a positive effect both on the original dependent variable and on its symmetric version.

The interpretation of the Irpino-Verde model is different as it takes into consideration the two components of the NPSD: the variability of the means and the variability of the centered NPSD (the variability of the distributions variability). The estimated model enounces that a unitary variation of the mean of the Hematocrit induces a variation of 3.146 in the mean of the Hemoglobin, while an increasing of one in the variability of the Hematocrit produce an increase of 3.969, in average, of the variability of the Hematocrit.

The second dataset derives from the Clean Air Status and Trends Network (CASTNET)⁴, an air quality monitoring network of United States designed to provide data to assess trends in air quality, atmospheric deposition, and ecological effects due to changes in air pollutant emissions. In particular, we have chosen to select data on the Ozone concentration in 78 USA sites among those depicted in Fig. 2 for which the monitored data was complete (i.e. without missing values for each of the selected characteristics).

Ozone is a gas that can cause respiratory diseases. In the literature there exists studies that relate the Ozone concentration level to the Temperature, the Wind speed and the Solar radiation (see for example [28]).

Given the distribution of Temperature (X_1) (Celsius degrees), the distribution of Solar Radiation (X_2) (Watts per square meter) and the distribution of Wind Speed (X_3) (meters per second),

⁴<http://java.epa.gov/castnet/>

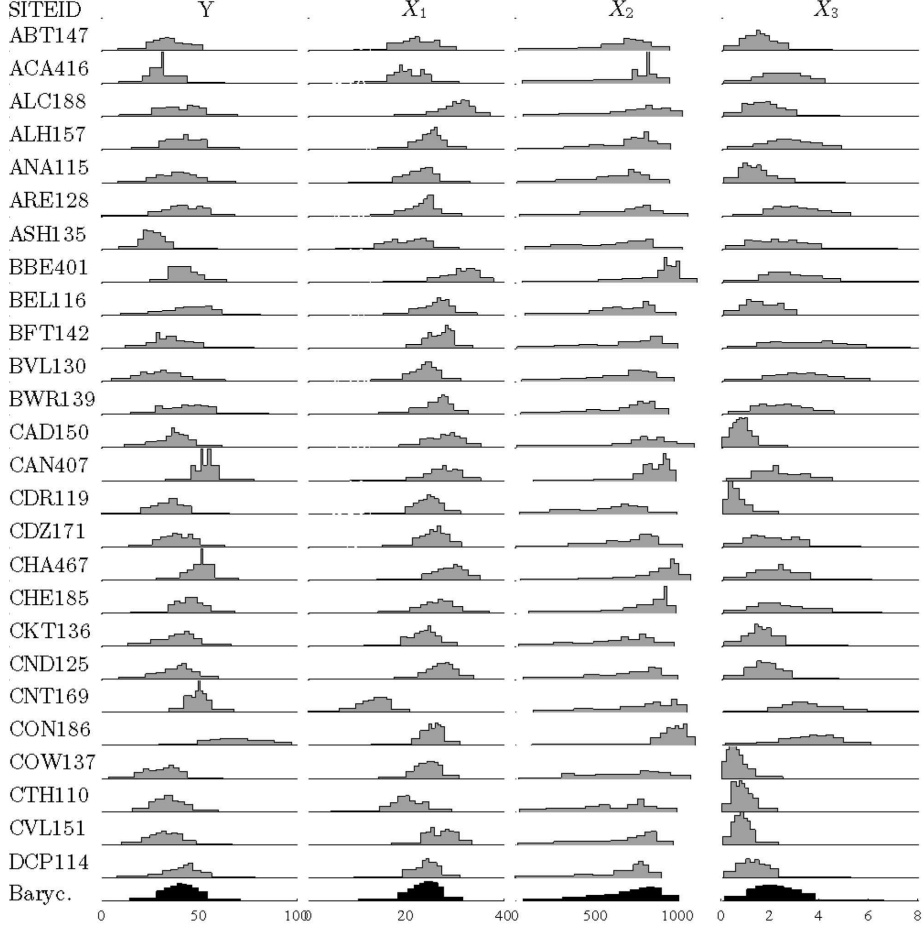


Figure 3: Ozone dataset: monitored sites from 1 to 26.

the main objective is to predict the distribution of Ozone Concentration (Y) (Particles per billion) using a linear model. CASTNET collect hourly data and as period of observation we choose the summer seasons of 2010 and the central hours of the days (10 a.m. - 5 p.m.).

For each sites we have collected the NPSD of the four variables in terms of histograms and in figs. 3, 4 and 5 we present the monitored sites using their histograms⁵, in order to have a reference we report in the last rows the barycenters that are better shown in fig. 6.

In table 6 we reported the main summary statistics for the four histogram variables, while in figure 6 are drawn the four barycenters where it is possible to observe the average distributions (in the sense of Verde and Irpino [26]) of the 78 sites for each variable.

We can note, for example, the different skewness of the barycenters, in general when the barycenter is skew, we may observe that the NPSD are in general skew in the same direction. This is not in general true for symmetric barycenters, that can be generated both from left and

⁵We can supply the full table of histogram data, the Matlab routines and the workspaces upon request.

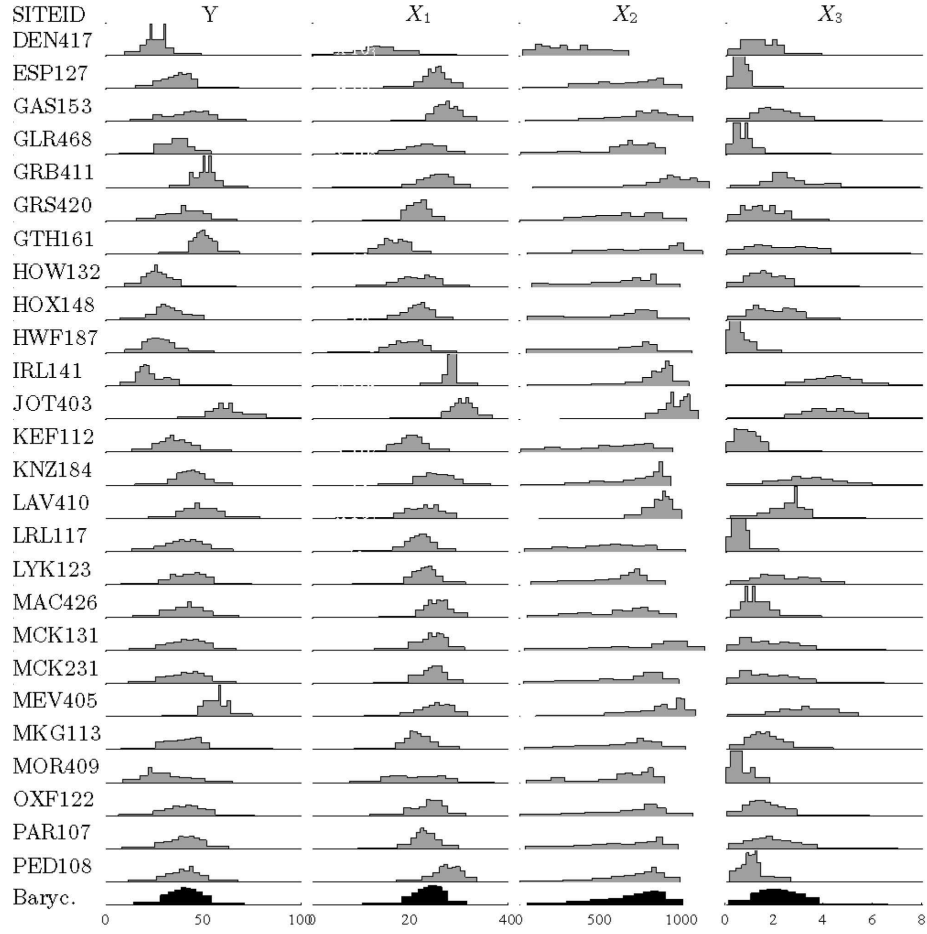


Figure 4: Ozone dataset: monitored sites from 27 to 52.

	Ozone Concentration (Y in Ppb)	Temperature (X_1 in Celsius deg.)	Solar Radiation (X_2 Watt/ m^2)	Wind Speed (X_3 m/s)		
Mean (BD)	41.2147	23.2805	645.3507	2.3488		
Barycenter mean (VI)	41.2147	23.2805	645.3507	2.3488		
Barycenter std (VI)	9.9680	3.7641	225.7818	1.0987		
Standard dev. (BD)	13.790	5.3787	252.6736	1.7125		
Standard dev. (VI)	9.5295	3.8422	113.4308	1.1337		
Correlations						
	Billard-Diday			Verde-Irpino		
	X_1	X_2	X_3	X_1	X_2	X_3
Y	0.2328	0.4064	0.2951	0.2473	0.6392	0.4020
X_1		0.2622	0.0621		0.4537	0.1429
X_2			0.3013			0.4394

Table 6: Ozone dataset: summary statistics

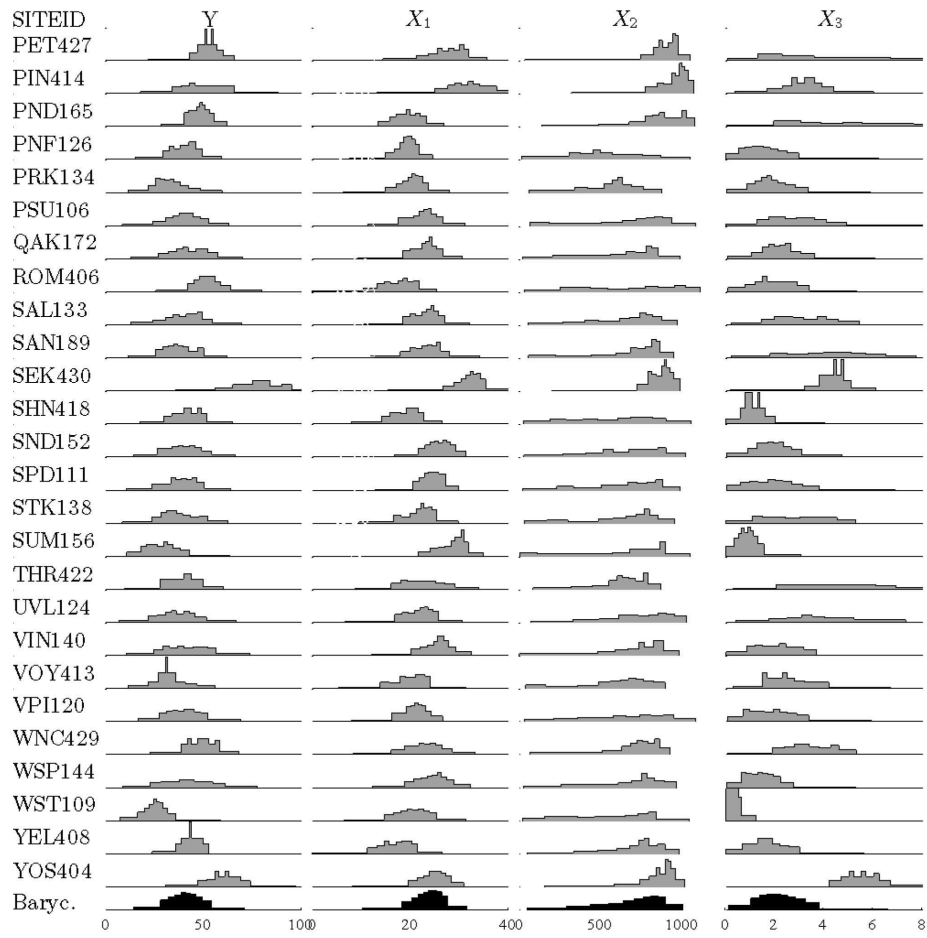


Figure 5: Ozone dataset: monitored sites from 53 to 78.

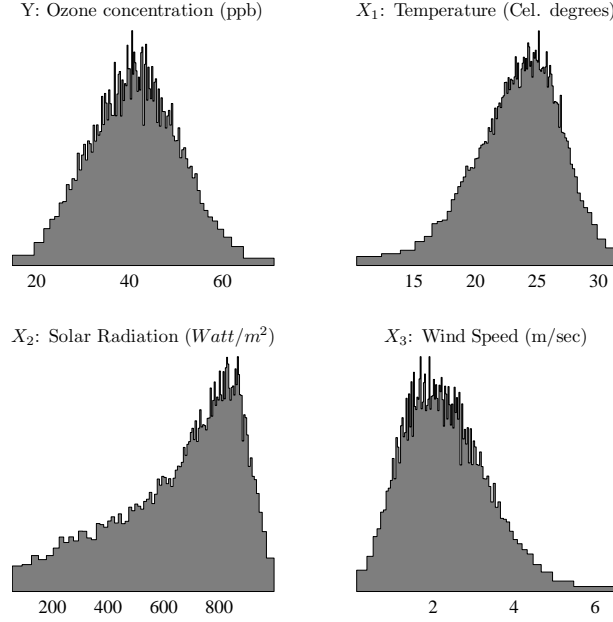


Figure 6: Ozone dataset. Barycenters

right skewed distributions.

Using the full dataset we estimated the three models and the associated goodness of fit diagnostics as reported in tables 7, 8 and 9.

Also in this case, we performed a bootstrap estimates of the pentameters and of the goodness of fit measures of the three models in tables 7, 8 and 9.

Observing the goodness of fit measures of the three models we can conclude that the Irpino-Verde and the Dias-Brito model fit better the linear regression relationship than the Billard-Diday model, and the Irpino-Verde model is slightly more accurate than the Dias-Brito one. Also in this case, the Irpino-Verde model parameters give an easier interpretation. Reading the Irpino-Verde bootstrapped model, we may assert that the Ozone concentration distribution of a site depends from the mean *solar radiation* where for each $\Delta \text{Watt/m}^2$ a $0.070(ppb)$ variation of the *Ozone concentration* mean level it is expected, while in general we cannot say that the mean level of *temperature* and of *wind speed* induces a significant variation of the *ozone concentration level* (the 95% bootstrap confidence intervals include zero). Furthermore, we may say that the variability of the *ozone concentration* is quite the same as the *temperature* (0.928), a unit variation in the variability of the *solar radiation* induces a variation of $0.018(ppb)$ and a variation of the variability of the *Wind Speed* causes an increase in the variability of $1.958(ppb)$. Similar conclusions can be derived reading the Dias-Brito model even if it gives a different interpretation of the parameter associated with the symmetric histogram variables.

Billard-Diday Model:							
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$							
	Model parameters				Goodness of fit		
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	Ω	$Pseudo - R^2$	$RMS E_W$
Observed	18.28	0.357	0.017	1.550	0.203	0.024	83.13
Bootstrap estimates							
Mean	18.96	0.323	0.018	1.463	0.215	0.103	81.91
Bias	0.678	-0.034	0.001	-0.087			
SE	5.077	0.182	0.005	0.652			
2.5%	9.52	-0.014	0.007	0.147	0.070	0.000	68.59
97.5%	28.89	0.697	0.027	2.836	0.350	0.372	96.85

Table 7: Ozone dataset: Billard-Diday model parameters estimated on the full dataset and bootstrapping the dataset.

Dias-Brito model:										
$\hat{y}_i(t) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}(t) + \hat{\beta}_2 x_{i2}(t) + \hat{\beta}_3 x_{i3}(t) + \hat{\beta}_1 \tilde{x}_{i1}(t) + \hat{\beta}_2 \tilde{x}_{i2}(t) + \hat{\beta}_3 \tilde{x}_{i3}(t)$										
	Model parameters							Goodness of fit		
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	Ω	$Pseudo - R^2$	$RMS E_W$
Observed	13.32	0.000	0.037	1.691	0.000	0.000	0.000	0.670	0.371	66.74
Bootstrap estimates										
Mean	14.22	0.117	0.034	1.709	0.080	0.000	0.002	0.712	0.358	65.07
Bias	0.905	0.117	-0.003	0.018	0.080	0.000	0.002			
SE	4.760	0.161	0.004	0.610	0.112	0.000	0.025			
2.5%	5.409	0.000	0.026	0.602	0.000	0.000	0.000	0.625	0.220	49.58
97.5%	24.46	0.540	0.040	3.070	0.391	0.000	0.000	0.801	0.498	80.60

Table 8: Ozone dataset: Dias-Brito model parameters estimated on the full dataset and bootstrapping the dataset.

Irpino-Verde model:										
$\hat{y}_i(t) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_{i1} + \hat{\beta}_2 \bar{x}_{i2} + \hat{\beta}_3 \bar{x}_{i3} + \hat{\gamma}_1 x_{i1}^c(t) + \hat{\gamma}_2 x_{i2}^c(t) + \hat{\gamma}_3 x_{i3}^c(t)$										
	Model parameters							Goodness of fit		
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	Ω	$Pseudo - R^2$	$RMS E_W$
Observed	2.928	-0.346	0.070	0.395	0.915	0.018	1.887	0.742	0.460	61.82
Bootstrap estimates										
Mean	3.108	-0.353	0.070	0.363	0.928	0.018	1.958	0.758	0.474	59.43
Bias	0.181	-0.008	0.000	-0.032	0.013	0.000	0.071			
SE	7.180	0.271	0.010	0.823	0.237	0.003	0.542			
2.5%	-11.24	-0.846	0.052	-1.186	0.482	0.012	1.054	0.675	0.296	46.52
97.5%	18.87	0.173	0.090	2.030	1.377	0.024	3.115	0.829	0.625	73.29

Table 9: Ozone dataset: Irpino-Verde model parameters estimated on the full dataset and bootstrapping the dataset.

6. Conclusions

The paper present a novel linear regression technique for data described by probability-like distributions, using their quantile functions and the ordinary least squares method based on the Wasserstein distance. Considering the nature of the data we proposed to use a particular decomposition of the Wasserstein distance for the definition of the regression model. We have also furnished an alternative goodness of fit index which takes into account the differences in shape and size of the quantile distributions of the independent variables. The proposed model corroborated on the above examples seems to have a better performances, in terms of goodness of fit, with respect the two main approaches presented in the literature. Further it allows an easier interpretation of the results. We also showed that the method can be used with a variety of numerical probabilistic symbolic data. Considering the complexity of the error term, the classic parameter inferential properties can not straightforward be extended to the regression of NPSD. We consider to address new efforts in the direction of investigate the properties of the involved estimators.

Appendix A. OLS solution details

Here we illustrate the OLS main passages

$$\begin{aligned}
 SS(\mathbf{B}, \mathbf{\Gamma}) &= \mathbf{e}^T \mathbf{e} = \\
 &= [\mathbf{Y} - \bar{\mathbf{X}}_+ \mathbf{B} - \mathbf{X}^c \mathbf{\Gamma}]^T [\mathbf{Y} - \bar{\mathbf{X}}_+ \mathbf{B} - \mathbf{X}^c \mathbf{\Gamma}] = \\
 &= \mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \bar{\mathbf{X}}_+ \mathbf{B} - \mathbf{Y}^T \mathbf{X}^c \mathbf{\Gamma} - \mathbf{B}^T \bar{\mathbf{X}}_+^T \mathbf{Y} + \mathbf{B}^T \bar{\mathbf{X}}_+^T + \bar{\mathbf{X}}_+ \mathbf{B} + \\
 &+ \mathbf{B}^T \bar{\mathbf{X}}_+^T \mathbf{X}^c \mathbf{\Gamma} - \mathbf{\Gamma}^T \mathbf{X}^{cT} \mathbf{Y} + \mathbf{\Gamma}^T \mathbf{X}^{cT} \bar{\mathbf{X}}_+ \mathbf{B} + \mathbf{\Gamma}^T \mathbf{X}^{cT} \mathbf{X}^c \mathbf{\Gamma} = \\
 &= \mathbf{Y}^T \mathbf{Y} - 2\mathbf{B}^T \bar{\mathbf{X}}_+^T \mathbf{Y} + \mathbf{B}^T \bar{\mathbf{X}}_+^T \bar{\mathbf{X}}_+ \mathbf{B} + 2\mathbf{B}^T \bar{\mathbf{X}}_+^T \mathbf{X}^c \mathbf{\Gamma} + \\
 &- 2\mathbf{\Gamma}^T \mathbf{X}^{cT} \mathbf{Y} + \mathbf{\Gamma}^T \mathbf{X}^{cT} \mathbf{X}^c \mathbf{\Gamma}
 \end{aligned} \tag{A.1}$$

first order conditions

$$\begin{aligned}
 \frac{\partial SS}{\partial \mathbf{B}} &= -2\bar{\mathbf{X}}_+^T \mathbf{Y} + 2\bar{\mathbf{X}}_+^T \bar{\mathbf{X}}_+ \mathbf{B} + 2\bar{\mathbf{X}}_+^T \mathbf{X}^c \mathbf{\Gamma} = 0 \\
 \frac{\partial SS}{\partial \mathbf{\Gamma}} &= +2\mathbf{B}^T \bar{\mathbf{X}}_+^T \mathbf{X}^c - 2\mathbf{X}^{cT} \mathbf{Y} + 2\mathbf{X}^{cT} \mathbf{X}^c \mathbf{\Gamma} = 0 \\
 \frac{\partial SS}{\partial \mathbf{B}} = 0 &\rightarrow -2 \underbrace{\bar{\mathbf{X}}_+^T \mathbf{Y}}_{\bar{\mathbf{X}}_+^T \bar{\mathbf{Y}}} + 2\bar{\mathbf{X}}_+^T \bar{\mathbf{X}}_+ \mathbf{B} + 2 \underbrace{\bar{\mathbf{X}}_+^T \mathbf{X}^c \mathbf{\Gamma}}_0 = 0 \rightarrow \mathbf{B} = (\bar{\mathbf{X}}_+^T \bar{\mathbf{X}}_+)^{-1} \bar{\mathbf{X}}_+^T \bar{\mathbf{Y}} \\
 \frac{\partial SS}{\partial \mathbf{\Gamma}} = 0 &\rightarrow 2\mathbf{B}^T \underbrace{\bar{\mathbf{X}}_+^T \mathbf{X}^c}_0 - 2 \underbrace{\mathbf{X}^{cT} \mathbf{Y}}_{\mathbf{X}^{cT} \mathbf{Y}^c} + 2\mathbf{X}^{cT} \mathbf{X}^c \mathbf{\Gamma} = 0 \rightarrow \mathbf{\Gamma} = (\mathbf{X}^{cT} \mathbf{X}^c)^{-1} \mathbf{X}^{cT} \mathbf{Y}^c
 \end{aligned}$$

Appendix B. Sum of square error decomposition

Considering $\mathbf{1} = [1]_{n \times 1}$, SSY can be written as:

$$\begin{aligned}
 SSY &= n \cdot s_y^2 = \sum_{i=1}^n d_W^2(y_i(t), \bar{y}(t)) = \sum_{i=1}^n \int_0^1 [y_i(t) - \bar{y}(t)]^2 dt = \\
 &= \left(\underbrace{\bar{\mathbf{Y}} + \mathbf{Y}^c}_{\mathbf{Y}} - \mathbf{1} \underbrace{(\bar{y} + \bar{y}^c(t))}_{\bar{y}(t)} \right)^T \left(\underbrace{\bar{\mathbf{Y}} + \mathbf{Y}^c}_{\mathbf{Y}} - \mathbf{1} \underbrace{(\bar{y} + \bar{y}^c(t))}_{\bar{y}(t)} \right)
 \end{aligned}$$

further

$$\begin{aligned}
SSY &= \int_0^1 (\mathbf{Y} - \mathbf{1}\bar{y}(t) + \hat{\mathbf{Y}} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \mathbf{1}\bar{y}(t) + \hat{\mathbf{Y}} - \hat{\mathbf{Y}}) dt = \\
&= \int_0^1 \left(\underbrace{\mathbf{Y} - \hat{\mathbf{Y}}}_{\varepsilon(t)} - (\mathbf{1}\bar{y}(t) - \hat{\mathbf{Y}}(t)) \right)^T \left(\underbrace{\mathbf{Y}(t) - \hat{\mathbf{Y}}(t)}_{\varepsilon(t)} - (\mathbf{1}\bar{y}(t) - \hat{\mathbf{Y}}(t)) \right) dt = \\
&= \int_0^1 \left(\underbrace{\varepsilon(t)^T \varepsilon(t)}_{SSE} + \underbrace{(\mathbf{1}\bar{y}(t) - \hat{\mathbf{Y}}(t))^T (\mathbf{1}\bar{y}(t) - \hat{\mathbf{Y}}(t))}_{SSR} - 2(\mathbf{1}\bar{y}(t) - \hat{\mathbf{Y}}(t))^T \varepsilon(t) \right) dt = \\
&= SSE + SSR - 2 \int_0^1 (\mathbf{1}\bar{y}(t) - \hat{\mathbf{Y}}(t))^T \varepsilon(t) dt = \\
&= SSE + SSR - 2 \left[\underbrace{\int_0^1 (\mathbf{1}\bar{y}(t))^T \varepsilon(t) dt}_{(I)} - \underbrace{\int_0^1 (\hat{\mathbf{Y}}(t))^T \varepsilon(t) dt}_{(II)} \right] =
\end{aligned}$$

for (I) we have

$$\begin{aligned}
\underbrace{\int_0^1 (\mathbf{1}_{n \times 1} \bar{y}(t))^T \varepsilon(t) dt}_{(I)} &= \int_0^1 (\mathbf{1}\bar{y}(t))^T (\mathbf{Y}(t) - \bar{\mathbf{X}}_+ \mathbf{B} - \mathbf{X}^c(t) \mathbf{\Gamma}) dt = \\
&= \int_0^1 (\bar{y}(t))^T \mathbf{1}^T \mathbf{Y}(t) dt - \int_0^1 (\bar{y}(t))^T \mathbf{1}^T \bar{\mathbf{X}}_+ \mathbf{B} dt - \int_0^1 (\bar{y}(t))^T \mathbf{1}^T \mathbf{X}^c(t) \mathbf{\Gamma} dt = \\
&= \int_0^1 (\bar{y}(t))^T n \cdot \bar{y}(t) dt - n\bar{y}^2 - \int_0^1 (\bar{y}(t))^T n \bar{\mathbf{X}}^c(t) \mathbf{\Gamma} dt = n \cdot \sigma_{\bar{y}}^2 + n\bar{y}^2 - n\bar{y}^2 - n \sum_{i=1}^p \gamma_p r_{\bar{y}\bar{x}_j} \sigma_{\bar{y}} \sigma_{\bar{x}_j} = \\
&= n \cdot \sigma_{\bar{y}}^2 - n \sum_{i=1}^p \gamma_p r_{\bar{y}\bar{x}_j} \sigma_{\bar{y}} \sigma_{\bar{x}_j} = n \cdot \left(\sigma_{\bar{y}}^2 - \sum_{i=1}^p \gamma_p r_{\bar{y}\bar{x}_j} \sigma_{\bar{y}} \sigma_{\bar{x}_j} \right)
\end{aligned}$$

for (II) we have

$$\begin{aligned}
& - \underbrace{\int_0^1 \left(\hat{\mathbf{Y}}(t) \right)^T \varepsilon(t) dt}_{(II)} = - \int_0^1 \left(\tilde{\mathbf{X}}_+ \mathbf{B} + \mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\mathbf{Y}(t) - \tilde{\mathbf{X}}_+ \mathbf{B} - \mathbf{X}^c(t) \mathbf{\Gamma} \right) dt = \\
& = - \int_0^1 \left(\tilde{\mathbf{X}}_+ \mathbf{B} \right)^T \mathbf{Y}(t) dt + \int_0^1 \left(\tilde{\mathbf{X}}_+ \mathbf{B} \right)^T \tilde{\mathbf{X}}_+ \mathbf{B} dt + \int_0^1 \left(\tilde{\mathbf{X}}_+ \mathbf{B} \right)^T \mathbf{X}^c(t) \mathbf{\Gamma} dt + \\
& - \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\mathbf{Y}(t) \right) dt + \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\tilde{\mathbf{X}} \beta \right) dt + \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right) dt = \\
& = \underbrace{- \tilde{\mathbf{Y}} \left(\tilde{\mathbf{X}} \beta \right)^T + \left(\tilde{\mathbf{X}} \beta \right)^T \left(\tilde{\mathbf{X}} \beta \right)}_{\beta \nabla f(\beta)=0} + \underbrace{\left(\tilde{\mathbf{X}} \beta \right)^T \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right) dt}_{0} + \\
& - \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\mathbf{Y}(t) \right) dt + \underbrace{\int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\tilde{\mathbf{X}} \beta \right) dt}_{0} + \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right) dt = \\
& = - \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\mathbf{Y}(t) \right) dt + \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right) dt = \\
& = - \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\tilde{\mathbf{Y}} + \mathbf{Y}^c(t) \right) dt + \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right) dt = \\
& = - \underbrace{\int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \tilde{\mathbf{Y}} dt}_{0} - \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \mathbf{Y}^c(t) dt + \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right) dt = \\
& = \underbrace{\int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right) dt}_{\text{Oif using OLS}} - \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \mathbf{Y}^c(t) dt = \mathbf{\Gamma} \nabla f(\mathbf{\Gamma}) \\
& \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right) dt - \int_0^1 \left(\mathbf{X}^c(t) \mathbf{\Gamma} \right)^T \mathbf{Y}^c(t) dt = \\
& = \mathbf{\Gamma} \underbrace{\int_0^1 \left[\left(\mathbf{X}^c(t) \right)^T \mathbf{X}^c(t) \mathbf{\Gamma} - \left(\mathbf{X}^c(t) \right)^T \mathbf{Y}^c(t) \right] dt}_{\nabla f(\mathbf{\Gamma})}
\end{aligned}$$

thus

$$bias = -2 \left[n \cdot \left(\sigma_{\tilde{\mathbf{y}}}^2 - \sum_{j=1}^p \gamma_j r_{\tilde{\mathbf{y}} \tilde{\mathbf{x}}_j} \sigma_{\tilde{\mathbf{y}}} \sigma_{\tilde{\mathbf{x}}_j} \right) + \mathbf{\Gamma} \nabla f(\mathbf{\Gamma}) \right]$$

the decomposition is then:

$$SSY = SSE + SSR - 2 \left[n \cdot \left(\sigma_{\bar{y}}^2 - \sum_{j=1}^p \gamma_j r_{\bar{y}\bar{x}_j} \sigma_{\bar{y}} \sigma_{\bar{x}_j} \right) + \mathbf{\Gamma} \nabla f(\mathbf{\Gamma}) \right]$$

References

- [1] H. Bock, E. Diday, Analysis of symbolic data: exploratory methods for extracting statistical information from complex data, Springer verlag, 2000.
- [2] L. Billard, E. Diday, Symbolic data analysis: conceptual statistics and data mining, Wiley, 2006.
- [3] E. Diday, M. Noirhomme-Fraiture, Symbolic Data Analysis and the SODAS software, Wiley, 2008.
- [4] P. Bertrand, F. Goupil, Descriptive statistics for symbolic data, in: H.-H. Bock, E. Diday (Eds.), Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Springer Berlin Heidelberg, 103–124, 2000.
- [5] E. de A. Lima Neto, F. de A. T. de Carvalho, C. P. Tenorio, Univariate and Multivariate Linear Regression Methods to Predict Interval-Valued Features, in: Australian Conference on Artificial Intelligence, 526–537, 2004.
- [6] E. de A. Lima Neto, F. de A. T. de Carvalho, Centre and Range method for fitting a linear regression model to symbolic interval data, Computational Statistics & Data Analysis 52 (3) (2008) 1500–1515.
- [7] E. de A. Lima Neto, F. de A. T. de Carvalho, Constrained linear regression models for symbolic interval-valued variables, Computational Statistics & Data Analysis 54 (2) (2010) 333–347.
- [8] R. Verde, A. Irpino, Ordinary Least Squares for Histogram Data Based on Wasserstein Distance, in: Y. Lechevallier, G. Saporta (Eds.), Proceedings of COMPSTAT'2010, chap. 60, Physica-Verlag HD, Heidelberg, 581–588, 2010.
- [9] L. Wasserstein, Markov processes over denumerable products of spaces describing large systems of automata, Prob. Inf. Transmission 5 (1969) 47–52.
- [10] L. Kantorovich, On one effective method of solving certain classes of extremal problems, Dokl. Akad. Nauk 28 (1940) 212215.
- [11] C. Gini, Di una misura della dissomiglianza tra due gruppi di quantit  e delle sue applicazioni allo studio delle relazioni statistiche, Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti, Tomo LXXIV parte seconda.
- [12] T. Salvemini, Sul calcolo degli indici di concordanza tra due caratteri quantitativi, in: Atti della VI Riunione della Soc. Ital. di Statistica, Roma, 1943.
- [13] G. Dall'Aglio, Sugli estremi dei momenti delle funzioni di ripartizione doppia, Ann. Scuola Normale Superiore Di Pisa, Cl. Sci 3(1) (1956) 3374.
- [14] C. L. Mallows, A Note on Asymptotic Joint Normality, The Annals of Mathematical Statistics 43 (2) (1972) 508–515.
- [15] P. Bickel, D. Freedman, Some asymptotic theory for the bootstrap, Ann. Stat. 9 (1981) 1196–1217.
- [16] A. Irpino, R. Verde, Dynamic clustering of interval data using a Wasserstein-based distance, Pattern Recognition Letters 29 (11) (2008) 1648 – 1658.
- [17] A. Irpino, R. Verde, A New Wasserstein Based Distance for the Hierarchical Clustering of Histogram Symbolic Data, in: V. Batagelj, H.-H. Bock, A. Ferligoj, A. Žiberna (Eds.), Data Science and Classification, Studies in Classification, Data Analysis, and Knowledge Organization, chap. 20, Springer Berlin Heidelberg, 185–192, 2006.
- [18] R. Verde, A. Irpino, Dynamic Clustering of Histogram Data: Using the Right Metric, in: P. Brito, G. Cucumel, P. Bertrand, F. Carvalho (Eds.), Selected Contributions in Data Analysis and Classification, Studies in Classification, Data Analysis, and Knowledge Organization, chap. 12, Springer Berlin Heidelberg, Berlin, Heidelberg, 123–134, 2007.
- [19] S. Dias, P. Brito, A new linear regression model for histogram-valued variables, in: 58th ISI World Statistics Congress, Dublin, Ireland, URL <http://isi2011.congressplanner.eu/pdfs/950662.pdf>, 2011.
- [20] A. Irpino, E. Romano, Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation, in: M. Noirhomme-Fraiture, G. Venturini (Eds.), EGC, vol. RNTI-E-9 of *Revue des Nouvelles Technologies de l'Information*, Cepad  s-  ditions, 99–110, 2007.
- [21] M. Noirhomme-Fraiture, P. Brito, Far beyond the classical data models: symbolic data analysis, Statistical Analysis and Data Mining 4 (2) (2011) 157–170, ISSN 1932-1872, doi:10.1002/sam.10112.
- [22] A. Irpino, R. Verde, Multiple regression via OLS for histogram symbolic data, in: U. of Pavia (Ed.), Cladag 2011 conference, CLADAG, 2011.
- [23] J. A. Cuesta-Albertos, C. Matr  n, A. Tuero-D  az, Optimal transportation plans and convergence in distribution, J. Multivar. Anal. 60 (1997) 72–83.
- [24] A. Irpino, R. Verde, Y. Lechevallier, Dynamic clustering of histograms using Wasserstein metric, in: COMPSTAT, 869–876, 2006.

- [25] C. L. Lawson, R. J. Hanson, Solving Least Square Problems, Prentice Hall, Edgeworth Cliff, NJ, 1974.
- [26] R. Verde, A. Irpino, Comparing Histogram Data Using a Mahalanobis-Wasserstein Distance, in: P. Brito (Ed.), COMPSTAT 2008, chap. 7, Physica-Verlag HD, Heidelberg, 77–89, 2008.
- [27] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.
- [28] C. Dueas, M. Fernández, S. Caete, J. Carretero, E. Liger, Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean Coast, Science of The Total Environment 299 (1-3) (2002) 97 – 113.